

Index

- African American Language (AAL), 185, 209, 210
 Corpus of Regional African American Language (CORAAAL), 164–165, 185–188, 190–191
 Early Black English corpora, 210
- African American Vernacular English (AAVE), 209–210, 211
- Alaska Native Languages Center, 134
- Algonquian languages (Cree)
 child language acquisition, 392, 393–395, 397–398
 transcription and processing of, 393–395
- Altmetrics, 159–160, 162
 altmetrics manifesto, 160
 for data sets, 166
- American Sign Language (ASL), 367
 ASL Signbank, 369, 375–377
 Black ASL Project, 474
 frequency effects, 476, 477
 genre, 475, 477
 Historical Sign Language Database, 478
 ID Gloss Project, 369, 375
 ID glosses, 375, 376, 377–378
 Internet-based research, 476–478
 varieties, 474, 477
 written glosses used for, 368, 369, 375
- Annotation, 30–33, 75. *See also* ELAN (annotation tool)
 bilingual, 214, 374, 379, 392
 of code-switching, 214, 392, 428
 coding, 83, 216, 613
 of cognate sets, 348
 corpus linguistics, 456–458
 as data, 33, 83
 database building, 395–397, 525
 indexing of recorded data, 251–252
 inline versus stand-off, 33
 interannotator agreement, 83, 501, 606
 interrater reliability, 388, 598
 Leipzig Glossing Rules, The, 31, 619
 machine learning techniques, 464, 468, 501
 multimodal, 379, 465–467
 reliability, 387, 388, 493–494
 of signed language video data, 272, 369, 374–379, 465–467, 468
 of sociolinguistic variation, 212–214, 281
 structural analyses, 33–36
 tools for, 32–33, 225, 281–282, 296, 465–467, 611
 treebanks, 33, 499, 501
 using XML, 33
- Anonymization of data, 80, 85, 104–105, 110. *See also*
 Confidentiality; Personal data
 CA (Conversation Analysis) data, 262
 child video data, 398
 corpora speakers, 189, 211, 408
 sign language video data, 270–271
- Anthropological linguistics, 40
- Application data, citation of, 151
- Archival documentation, 315–317, 318
 creation of language and literacy resource materials, 323
 data management, 319–322, 322
 organization of data, 318, 321
 software, 321–322
- Archive of the Indigenous Languages of Latin America (AILLA), 92, 96, 134
- Archives. *See* Language archives
- Archiving data. *See also* Data repositories
 barriers to, 91–92, 518
 benefits of, 16, 90, 109, 263, 518
 digital tools for, 15–16, 19
 documentation projects, 288, 297, 331
 inheritance plans for archived data, 110
 metadata standards for, 41–42, 66–68, 94, 95–96, 110
 primary and secondary data, 200, 336
 rates of, 447
 readme files, 67–68, 108, 449
 requirements for, 67, 94, 95–96, 110
 for revitalization, 135, 137–138, 283
 scripts (computer), 447, 494, 532, 535
 selection of data for, 95, 110
 signed language video data, 379, 467, 476
 speech perception data, 570, 571
- Archiving research data. *See* Data repositories: selection of
- ASL (American Sign Language), 367
 ASL Signbank, 369, 375–377
 Black ASL Project, 474
 frequency effects, 476, 477
 genre, 475, 477

- ASL (American Sign Language) (cont.)
 Historical Sign Language Database, 478
 ID Gloss Project, 369, 375
 ID glosses, 375, 376, 377–378
 Internet-based research, 476–478
 varieties, 474, 477
 written glosses used for, 368, 369, 375
- Atlas of North American English, 237
- Audio British National Corpus (AudioBNC), 197, 198
- Audio recordings
 anonymization of, 189, 262
 collection of, 250, 280–281, 328, 403
 decision not to record, 514
 file-naming templates, 441
 informed consent, 187–188, 328 (*see also* Consent)
 metadata, 251, 291, 328, 331
 phone conversations, 250
 processing and storage of, 251, 291
 quality of, 440
 recording equipment, 239, 250, 333, 403
 signadome (mobile recording studio), 440
 storage of data, 240, 281, 291, 298
 time codes in field notes, 291
 wax cylinders, 137, 316
- Austin Principles of Data Citation in Linguistics, The, 5, 51, 144
- Authorship
 coauthorship, 166, 199–200
 of data for citation purposes, 146, 148, 149
 legal definition of, 118
- Automatic speech recognition (ASR), 195–196, 438, 440
 data quality, 440
 signed language recognition, 463–464
 speech perception data, 565, 568
- AUTOTYP (typological database), 631–634, 634–638, 638–640
- Availability of data, 15, 144, 151. *See also* Data accessibility
- Bantu linguistics, 386, 609
- Bayesian phylogenetic methods, 349–351
- BibLaTeX (reference management tool), 152
- BibTeX (reference management tool), 152, 590, 635–636
- BISINDO corpus (Indonesian Sign Language), 268, 272
- Breath of Life, 137
- British National Corpus, 29, 453, 454
 Audio British National Corpus (AudioBNC), 197, 198
 XML edition, 453
- Buckeye Corpus, 197
- Byte Order Marks (BOMs), 77
- CA (Conversation Analysis) research, 257, 261–262
 classic data of, 258, 263
 data citation, 262
 data collection, 259
 data sharing in, 263
 transcription, 260–261
- California Language Archive, 134
- CESSDA Data Catalogue (Consortium of European Social Science Data Archives), 92
- Character encoding, 73–74, 77–78, 560
 repository data, 111
- CHAT transcription (Codes for the Human Analysis of Transcripts), 370, 387, 406–407
 bilingual signed and spoken data, 374
 orthography required by, 406–407
- Cherry-picking, 17
 minimizing for Internet data, 474
- CHILDES (Child Language Data Exchange System), 92, 370, 386, 387, 398
- Child language acquisition data, 369, 385–386, 392–393
 anonymization and deidentification of, 398
- CHAT transcription (Codes for the Human Analysis of Transcripts), 370, 387
- Chisasibi Child Language Acquisition Study (CCLAS), 392–393
- confidentiality of data, 380, 398
- corpora, 385–389, 392–393
- Databrary database, 388
- Indigenous languages, 392, 393–395, 397–398
- metadata, 374, 379
- multilingual contexts, 369, 392
- PhonBank, 388, 391, 392, 398
- sharing of, 379–380, 397–399
- signed language, 369, 371–373
- video data collection, 370–371, 387–388, 388, 393
- Chirila database of contemporary and historical lexical data for Australian languages, 335–338, 340
- Citation. *See* Data citation
- Citation-based metrics, 159, 160, 165–166
- CLARIN (Common Language Resources and Technology Infrastructure) Virtual Language Observatory, 93
- Clinical initiatives, language resources for, 398
- Cloud-based storage, 56, 66, 106, 298, 563
- Coding. *See* Annotation
- Coding scripts. *See* Computer scripts
- Cognitive linguistics
 construct validity, 493
 data sharing, 494
 interrater reliability, 493
 reproducibility, 491, 492–494
 sensory and perceptual language, 490
- Collaboration
 contracts, 123, 126
 international, 199, 273
 research assistants, 253, 403–404, 520
 teamwork, 303, 426–427, 523, 540, 563
 tools for, 84, 540, 563
 workflow, 253
- Collaborative consultation process, 53
- Colonialism, 221–222, 318, 323
- Command-line tools, 78, 83, 196

- Common Language Resources and Technology Infrastructure (CLARIN) Virtual Language Observatory, 93
- Community-based research, 53–54, 271, 278, 392, 397–399
 appraisal of data sets, 63
 community ownership of language, 50–52
 language and literacy resource materials, 279–280, 283, 301, 323
 public data access, 428, 429–430
 as requirement for language research, 53–54
 revitalization efforts, 318, 429
 standards for, 55–56
- Companion course (open access online), 6–7
- Comparative research. *See also* Historical linguistics; Language change
 coding schemas, 623–624
 comparative method, 345, 348, 349
 data collection, 610
 reclamation-driven reconstruction, 322
 theoretical linguistics, 622, 627
 tools for, 335–338, 339–340, 617
 underdescribed languages, 609, 611
- Compatibility of files
 cross-platform, 77–78
- Component Metadata Infrastructure (CMDI) framework, 42
- Computational modeling
 data management, 358, 360, 362, 544
 documentation of process, 358–359, 361–363
 exposition file, 360, 362, 364
 goals of, 356, 357
 information processing, 358
 of language evolution, 355–357, 359–360
 pseudo-code, 362–363
 simulation code clarity, 360, 362
 treebank data, 499
- Computational neurolinguistics, 547–548
 brain mapping, 548, 549, 550
 Canonical Correlation Analysis (CCorA), 550, 551
 fMRI (functional magnetic resonance imaging), 30, 547–548, 549
 semantic and neural networks, 549, 551
- Computational phylogenetics, 345
 Bayesian phylogenetic methods, 349–351
 Cross-Linguistic Data Formats (CLDF) initiative, 347–348
 data analysis, 348–351
 data format for, 347
 data sharing and reuse, 351
 tools for, 348, 349
- Computer scripts. *See also* Programming languages
 documentation of, 361–363, 447–451, 457–458, 535, 545, 561–562
 open source code, 432, 447, 450–451
 practices for coding, 456, 535, 545, 561–562
 pseudo-code, 362–363
 Python script for LaTeX example formatting, 525
 sharing of, 447, 494, 532, 535, 585, 639–640
- Concepticon initiative, 347
- Conceptual replications, 12
- Concordance software, 215
- Confidential data, 106, 110
- Confidentiality, 96–97, 103, 106, 110. *See also* Ethical practices;
 Personal data
 audio/video data, 252–253, 259
 signed language video data, 380, 472
- Consent, 104–105
 CA (Conversation Analysis) recordings, 259
 corpus data, 187–188, 217, 269–270, 402–403, 464–465
 of parent or guardian, 402–403
 reconsenting protocols, 370
 recorded orally, 269, 328
 regarding sharing of research data, 123, 254, 403
 signed language video data, 269–270, 369, 464–465, 472
- Construct validity, 493, 598
- Conversation Analysis (CA) research, 257, 261–262
 classic data of, 258, 263
 data citation, 262
 data collection, 259
 data sharing in, 263
 transcription, 260–261
- Copyright. *See* Data sets: copyright of
- Copyright law, 118–120
 all rights reserved, 124
 authorship, 118–119, 123
 Berne Convention for the Protection of Literary and Artistic Works, 120
 CC (Creative Commons) licenses, 111, 124, 464, 485, 501
 contracts and transfer agreements, 122, 198
 differences between countries, 119, 120, 121
 ethics of, 123–124
 Fair Use and Fair Dealing, 121–122, 125
 GNU General Public License (GNU GPL), 125
 Intellectual property (IP), 104, 118, 120, 124, 428
 Internet video ownership, 475
 open licenses, 124–125
 public domain, 119, 124
sui generis rights, 119
 Terms of Service and licensing agreements, 56, 122, 124–126, 485
 terms of use, 476
 when copyright applies, 120
- Corpora
 associative concept dictionaries, 549
 Audio British National Corpus (AudioBNC), 197, 198
 audio/video, 387, 388, 389, 427
 bilingual, 214, 369, 392
 British National Corpus, 29, 453, 454
 child language acquisition, 369, 385–386, 392–393
 CHILDES (Child Language Data Exchange System), 92, 370, 386, 387, 398
 citation of, 149, 217
 collection in school setting, 402–405, 407–408

- Corpora (cont.)
- collection of data for, 186, 210, 223, 386, 387–388, 388, 401
 - Corpus of Contemporary American English, 29, 489–490
 - Corpus of Regional African American Language (CORAAAL), 164–165, 185–188, 190–191
 - cross-corpus analysis, 195–197, 200, 206, 223, 413
 - curation of, 29
 - demographic makeup of, 186–188, 223, 404
 - Early Black English corpora, 210
 - errors in, 191, 214
 - file naming, 189
 - informed consent for, 187–188, 217, 269–270, 402–403, 464–465
 - Internet data, 485
 - Irvine Phonotactic Online Dictionary of English (IPhOD), 576
 - L2 learner corpora, 401, 406–407, 411, 412, 454
 - metadata, 189, 211, 464
 - natural language processing (NLP), 413–414
 - non-overt (deleted) forms in, 214, 215
 - open access corpora, 190, 197–198, 468
 - oral and written, 401, 403–404, 406
 - parallel corpora, 386, 388, 413
 - privacy and confidentiality, 189, 217, 269–270, 464–465
 - public corpora, 185, 188, 197
 - Santa Barbara Corpus of Spoken American English, 197, 198, 257
 - sharing of, 190, 217, 397–399, 406, 408
 - signed language corpora, 268, 463–468
 - sociolinguistic interviews, 185–186, 210–211, 216, 223
 - speaker selection for, 186, 209–210, 404
 - spoken language corpora, 195, 201, 209, 210
 - textual corpora, 28–29, 403, 406, 548
 - transcription, 188, 191, 225
 - treebanks (*see* Treebank data)
 - Twitter, 485
 - usefulness of, 385, 388, 389
 - XML edition of, 453
- Corpus linguistics, 28–29, 268, 453
- annotation, 456–458
 - annotation errors, 458, 459
 - child language corpora, 369, 385–386, 395–397
 - concordance generation, 454, 460
 - concordance software, 215
 - controlling for priming effects, 455, 457
 - corpus phonology, 575, 576, 582
 - corpus-based typology, 597
 - data management workflow, 454–456, 456–458, 459, 460
 - Korp (corpus querying tool), 611
 - metadata, 458, 460
 - programming languages, 454
 - reproducibility, 490
 - sampling corpus data, 457
 - semantic prosody, 489, 491
 - spreadsheet software, 458, 459
 - using R (software environment), 454–460
 - using the British National Corpus, 453
- Corpus neurolinguistics. *See* Computational neurolinguistics
- Corpus of Contemporary American English, 29, 489–490
- Corpus of Regional African American Language (CORAAAL), 164–165, 185–188, 190–191
- Corruption (data loss), 75, 84, 340
- Cree (Algonquian language)
- child language acquisition, 392, 393–395, 397–398
 - transcription and processing of, 393–395
- Crosslinguistic research, 597–598
- Cross-Linguistic Data Formats (CLDF) initiative, 347–348
 - Cross-Linguistic Linked Data (CLLD) project, 351
 - Cross-Linguistic Transcription Systems initiative, 347
 - tools for, 206
- CSV files, 74. *See also* File formats
- benefits of, 85, 415
 - limitations of, 84, 449
 - used for public archiving, 591
- Data. *See also* Data sets; Linguistic data
- accuracy of, 4
 - analog data, 102, 106
 - applied as term to Indigenous languages of North America, 316
 - cleaning of, 79–81, 85, 336, 611
 - corruption of, 75, 84, 340
 - definition of, 3, 102, 316
 - deidentification of, 270, 398, 408, 535
 - digital parameters of, 102
 - digitization of, 137, 138, 216, 387
 - documentation of, 66–68, 107–109, 338, 460, 558, 561 (*see also* Metadata)
 - errors in, 4, 81, 84, 282, 458, 625
 - ethics of (*see* Ethical considerations; Ethical practices)
 - experimental, 532
 - found materials as, 257
 - internal consistency of, 74, 458
 - legacy data, 51–52, 138
 - life cycle model of, 61–63, 341
 - linguistic (*see* Linguistic data)
 - multilingual, 392, 430, 611, 612, 613
 - naturally occurring, 28, 257
 - normalization of, 80, 159
 - as output of research, 4, 11, 16, 19, 90, 177–178
 - ownership of, 50–52, 93, 104
 - primary data, 50, 336
 - as primary source, 29, 90
 - raw data, 50, 82, 96, 106
 - representativeness of, 11–12, 28
 - reuse of, 61, 104, 124, 634
 - social sciences data, 10
 - storage and backup of, 66, 105, 251, 298, 329 (*see also* Data storage)
 - storage versus preservation, 110

- subjective data, 492
- underpinning publications, 5, 90, 145
- unstructured data, 610
- validation of, 80
- Data accessibility, 639–640. *See also* FAIR data principles
 - cited data excerpts, 297
 - Cross-Linguistic Linked Data (CLLD) project, 351
 - data accessibility statements, 145
 - Indigenous language data, 56, 428
 - language and literacy materials, 283, 341
 - principles of, 109, 125, 144
 - rewards of, 9, 90
 - signed language video data, 270, 377, 379–380, 476
 - unique identifiers for data, 448
- Data aggregation
 - aggregation pipeline, 589–591
 - late aggregation, 633–634, 638
 - metadata, 168, 589
- Data availability statements, 145, 149
- Databases
 - AUTOTYP (typological database), 631–634, 634–638, 638–640
 - Chirila database of contemporary and historical lexical data for Australian languages, 335–338, 340
 - Chisasibi Child Language Acquisition Study (CCLAS), 392–393, 395–397
 - citation of, 150, 626
 - curation of, 338, 341–342, 524, 620
 - Databrary database, 388
 - design of, 202, 341–342, 524, 619, 631–632, 633
 - distinguished from data (copyright), 119
 - ethical considerations for, 341
 - ImproType (Typological Database of Impersonals), 599–604, 604–606
 - Indigenous Tweets project, 483–485
 - IRIS repository for second language research, 408
 - metadata aggregation from, 589, 590
 - neurolinguistics, 549, 550–551
 - partial annotation of, 395
 - PHOIBLE (Phonetics Information Base and Lexicon Database), 589–591, 592
 - Phon (database software), 391, 395–397
 - PhonBank (child phonology database), 388, 391, 392, 398
 - regularization of primary data, 336, 338
 - relational database system, 605, 606, 631–632
 - reproducibility of decisions, 338
 - research strata in, 336
 - signbanks, 375–377, 467
 - signed languages, 369, 375–377, 478
 - software, 200, 339–340, 391, 481–482
 - TalkBank database system, 257, 391, 398–399
 - TerraLing (searchable database collection), 617–619, 620, 625–627
 - Twitter, 481–482
 - World Atlas of Language Structures (WALS), 618
- Data citation
 - application data, 151
 - Austin Principles of Data Citation in Linguistics, The, 5, 51, 144
 - bibliographic reference, 146, 148
 - of corpora, 199–200, 217
 - of corpora transcript and audio, 190
 - of corpus excerpts, 282
 - credit and attribution, 51, 90, 144, 146, 166, 627
 - data availability statements, 145, 149
 - examples of, 149–152, 262, 297, 626–627
 - footnotes or endnotes, 149
 - granularity of citation, 148
 - importance of, 16, 144, 164
 - Internet data, 476
 - in-text citation, 148, 263
 - Joint Declaration of Data Citation Principles (FORCE11), 143–144
 - metrics, 165–166
 - persistent identifiers for, 448, 507
 - principles of, 144, 145
 - publisher data policies and guidelines, 153–154
 - rates of, 10, 90, 447
 - recommendations for, 124, 145–146, 399, 566
 - of recordings, 31, 297
 - reference management tools, 152
 - of replication data sets, 151
 - Scholix initiative (to facilitate the exchange of data citations), 166
 - Science Citation Index, 157
 - signed language data, 476
 - source of linguistic examples, 31, 262, 297
 - standards for, 51
 - of subset of data (e.g., time span), 149, 190, 615
 - templates for, 146–147, 148
 - TerraLing data sets, 626–627
 - treebanks, 507
 - Tromsø Recommendations for Citation of Research Data in Linguistics, 5, 144
- DataCite Metadata Schema, 67, 146
- Data dictionaries, 68, 108
- Data documentation. *See* Metadata
- Data fixity, 147
- Data Guardians (DGs), 198, 200
- Data harvesting, 73
- Data journals, 163
- Data management. *See also* Data management plans (DMPs)
 - archival documentation, 319–322, 322
 - cleaning and normalization, 79–81, 228, 336
 - computational modeling, 358, 360
 - corpus linguistics, 454, 456–458, 459
 - databases, 336, 338, 341–342
 - documentary linguistics, 270, 278–283, 288, 327–331
 - ethical approaches to (*see* Ethical approaches to data management; Ethical practices)

- Data management (cont.)
 - experimental research, 535–536, 540, 541–542, 558, 561
 - experimental syntax, 531, 533–536
 - fieldwork, 269, 327–331, 333
 - file naming, 64, 65–66, 108–109, 216, 251
 - Internet data, 472, 476, 484–485
 - lexicography, 306, 307–309
 - life cycle model of, 61–63, 216, 341
 - organization of files, 78, 260, 540, 558
 - piloting, 401–402, 403, 542
 - pipelines and protocols, 340
 - postprocessing, 81–82, 85
 - principles of, 69, 83–84, 85–86, 94, 338
 - signed languages, 272–273, 367–368, 464–467, 472
 - sociolinguistic data, 186–190, 211–217, 223–225, 281
 - training and courses on Research Data Management (RDM), 91
 - transferring and updating files, 329
 - treebanks, 500, 503, 510
 - typology, 38–40, 597–600, 606, 632–633
 - untrained forced alignment, 424, 427–429, 429–432
 - value of, 4, 179, 245, 272, 288
 - workflow, 62, 64, 110, 112, 251, 327–331
- Data management plans (DMPs), 64, 101–102
 - archiving of, 111
 - confidentiality, 103, 106
 - good practices for, 64, 102, 105–106
 - guidelines for, 11
 - legal and ethical considerations, 104–105, 111–112
 - piloting, 401–402, 542
 - requirements of funders, 113
 - revision of, 112–113
 - tools and resources for, 101–102, 113
 - what to include, 102, 104, 105–106, 109, 111, 112
 - workflow for, 64, 101, 110, 112–113
- Data management use cases, online collection page, 7
- Data manipulation, 17
 - recommendations for preventing data inflation, 17–18
 - research fraud, 447
- Data mining of copyrighted material, 121–122
- Data papers, 173
- Data pipelines and bottlenecks, 340
- Data preservation, 110. *See also* Data storage
- Data protection, 105–107, 110–112
- Data quality, 159
 - acoustic analysis, 224
 - audio recordings, 440
 - metadata quality, 592
 - video recordings, 464, 465
- Data repositories. *See also* FAIR data principles; Language archives
 - access control, 96
 - Archive of the Indigenous Languages of Latin America (AILLA), 92, 96, 134
 - certification of, 93
 - CLARIN (Common Language Resources and Technology Infrastructure), 93
 - data licensing, 94, 95, 111, 124–126 (*see also* Copyright law)
 - Dataverse institutional repositories, 15, 92
 - Digital Endangered Languages and Musics Archives Network (DELAMAN), 93
 - domain-specific repositories, 92
 - Endangered Languages Archive, 136, 467
 - file formats, 64–65, 111, 448–449, 620, 640
 - file naming, 108
 - GitHub (repository hosting service), 15, 84, 540, 544
 - Indigenous Language Digital Archive, 322
 - inheritance plans for archived data, 110
 - IRIS database for second language research, 408
 - Kaipuleohone Language Archive, 134, 288, 297
 - OLAC (Open Language Archives Community), 93
 - Open Science Framework (OSF), 16, 536
 - Language Program, 136
 - Language Archive, The (TLA), 136, 467
 - metadata, 67–68, 94, 110, 153, 297, 448
 - PARADISEC (Pacific and Regional Archive for Digital Sources in Endangered Cultures), 92, 134, 331
 - project repository structure, 540
 - Registry of Research Data Repositories, 93
 - requirements for, 67, 94, 95–96, 110
 - role in data preservation, 51, 110, 163
 - selection of, 15–16, 92, 96, 110, 165, 519
 - selection of data for, 110
 - signed language data, 476
 - size limits, 544
 - Stanford Digital Repository, 519
 - treebanks, 507
 - Tromsø Repository of Language and Linguistics (TROLLing), 92, 94, 96, 448
- Data science, 73, 339
- Data security, 103, 105–107, 110–112
 - General Data Protection Regulation (GDPR), 94, 96, 198–199, 398
 - personal information, 96–97, 105, 123, 259 (*see also* Personal data)
 - repository access control, 96
 - video data, 252–253, 380, 472
- Data sets
 - augmentation of, 75
 - citation of, 5, 146, 151, 199–200, 448 (*see also* Data citation)
 - copyright of, 118, 123, 124–125 (*see also* Data sets: licensing of)
 - corruption of, 75, 84, 340
 - derived data sets, 196, 198
 - digital identifiers for, 147
 - discoverability of, 94, 96
 - embargoes on, 18, 96, 110
 - inheritance plans for, 110
 - interoperability of, 125, 144, 347, 448, 621
 - licensing of, 94, 95, 111, 124–126
 - metadata for, 79, 448
 - metrics for impact of, 165–166, 173
 - open access, 16, 125, 190, 197–198, 468, 565–566

- organization of files, 65–66, 74, 79, 108, 542
- peer review of, 96
- as product of research, 16, 90, 109, 173, 177–179
- publication date of, 147
- publication of, 15, 82, 92
- reorganization of files, 74, 78–79
- semantic norms, 490, 491–492
- sociolinguistic, 185–186, 209, 211
- sustainability of, 63
- titles of, 147, 166
- transliteration of, 74
- trimming or simplification of, 74
- working with multiple data sets, 125
- Data sharing
 - barriers to, 20, 91–92, 118
 - benefits of, 90, 144, 162, 341, 447–448, 518
 - best practices for journals and publishers, 168
 - blogging platform used for, 254
 - child language acquisition data, 379–380, 397–399
 - cognitive linguistics, 494, 550
 - considerations for, 20–21, 112, 125, 341, 408
 - context-dependent consent, 254
 - copyrighted works, 121–122
 - corpora, 190, 397–399, 468
 - of corpus collection methodology, 408
 - Cross-Linguistic Linked Data (CLLD) project, 351
 - data transfer agreements, 198
 - Dropbox used for, 379
 - ethics of, 20, 56, 125–126, 198–199, 253–254, 341
 - GitHub (repository hosting service), 540, 591
 - importance of, 173, 399, 406, 494
 - Indigenous language data, 397–399
 - maintaining closed data, 253–254
 - models of, 163, 199, 254, 263
 - model source codes, 362–364
 - policies on, 12
 - repository file protection, 93–94, 96
 - researcher-to-researcher, 263
 - R scripts, 450–451, 532, 534, 535
 - scripts (computer), 447, 494, 532, 585
 - signed language data, 273, 379–380, 468, 476
 - statistical code, 447, 449–450, 450–451
 - social media used for, 379
- Data sovereignty. *See* Indigenous data sovereignty (IDS)
- Data storage, 66, 105–106, 110. *See also* File formats
 - of audio recordings, 216, 240–241, 291
 - cloud storage, 56, 66, 106, 110, 298
 - Dropbox used for, 378
 - file formats, 64–65, 73–74
 - Git Large File Storage (Git-LFS), 544
 - rsync open source utility, 251
 - transferring and updating files, 329
 - treebanks, 502
 - versus data preservation, 74, 110
 - of video data, 372–374
- Data transformation
 - archival documentation, 319–322
 - tools for, 77, 79, 81, 83, 84, 425, 431
- Dataverse institutional repositories, 15, 92
- Date accessed (for citations), 147, 150
- Declaration on Research Assessment (DORA), 90, 158, 172
- Decolonization, 51, 315, 322–324
- Demographic data, 40
 - avoiding assumptions from video data, 474, 475
 - corpus participants, 186–188, 189, 211, 404
 - experimental syntax, 535
 - metadata, 189–190, 211, 223, 279, 615
- Demuth Sesotho Corpus (child language acquisition), 385, 386–387
- Derived data sets, 196, 198
- Description, linguistic. *See* Documentary linguistics
- Diachronic analysis, 35, 338
 - signed languages, 478
- Dialectology, 237
- Dictionaries, 37, 38
 - associative concept dictionaries, 549
 - bilingual or trilingual, 295, 302
 - considerations for creating, 301, 303, 304, 309–310
 - digital pipeline for creating, 306
 - ethnobiological terminology, 303
 - example sentences in, 311
 - FLEx used to create, 307–309
 - headword selection, 309–310
 - inclusion of cultural knowledge in, 303
 - as outcomes of language documentation, 301
 - structure of, 306
 - thematically organized, 303, 306, 309
 - tools for creating, 295, 306
 - workflow for creating, 307–309, 310, 311–312
- Digital archives, 15–16, 137. *See also* Language archives
- Digital Endangered Languages and Musics Archives Network (DELAMAN), 137
- Digital identifiers. *See* Identifiers
- Discoverability of data, 96, 162, 164–165, 566
- Dissemination of data, 109, 163, 536. *See also* Data sharing
 - community-based research, 397–399
- Dissemination of research, 104, 159, 161–162, 397–399. *See also* Publications
- DMPs (Data management plans), 64, 101–102
 - archiving of, 111
 - confidentiality, 103, 106
 - good practices for, 64, 102, 105–106
 - guidelines for, 11
 - legal and ethical considerations, 104–105, 111–112
 - piloting, 401–402, 542
 - requirements of funders, 113
 - revision of, 112–113
 - tools and resources for, 101–102, 113
 - what to include, 102, 104, 105–106, 109, 111, 112
 - workflow for, 64, 101, 110, 112–113

- Documentary linguistics
- archival documentation, 315–317, 318, 322
 - archives for, 134, 136 (*see also* Language archives)
 - collection of audio recordings, 280–281, 289, 291
 - community data access, 270, 428
 - corpus transcription, 427–429
 - creation of language and literacy resource materials, 279, 283, 302, 323, 429
 - creation of language dictionaries, 295, 301, 302, 310
 - data management practices, 277, 278, 282, 283, 291
 - data of, 28, 289
 - endangered languages, 136, 287 (*see also* Endangered language data)
 - equipment for, 330
 - ethics of, 52–54, 57, 271, 428–429, 432 (*see also* Ethical practices)
 - fieldwork, 288, 327–331
 - file naming, 281
 - GOLD ontology (General Ontology for Linguistic Description), 137
 - Indigenous languages, 49, 52–54, 57, 111–112, 428–429, 432 (*see also* Indigenous language data)
 - language work, 278
 - lexical data, 304, 310
 - local and distant documentation, 267–268
 - metadata, 279, 288
 - sharing of language and literacy resource materials, 283, 301, 429
 - sign languages, 267–268, 271, 468
 - tools for, 288, 296, 297
 - workflow, 282, 288, 331
- DOIs (digital object identifiers). *See also* Identifiers
- for data sets, 94, 448
 - recommendations for, 147
- Dropbox, used for syncing and backup, 378, 563
- Dynamic Variability in Speech (DyViS) forensic corpus, 197
- Education
- language and literacy resource materials, 279, 283, 323, 398, 429
 - role of data sets in, 91, 109
 - second language instruction, 401
- ELAN (annotation tool), 281, 287, 332
- and data citation, 282
 - file formats, 65, 296, 466–467
 - signed language video data, 374, 378–379, 465–467
 - types of annotations, 32–33
 - versus Transcriber, 225
 - workflow, 282–283, 296, 526
- Elicitation
- comparative analysis, 610, 625
 - of dictionary example sentences, 311
 - language description, 610, 625
 - language documentation, 288–289, 311, 328, 463, 473
 - lexical, 131, 304–306
 - to measure L2 proficiency, 403, 404, 405
 - narrative tasks, 465
 - of regional forms, 237, 239
 - signed language data, 472
 - tasks for, 132, 239, 289, 304–306, 404
 - theoretical syntax, 527–528
 - of vernacular variants, 210
- Elpis (speech recognition tool), 437–438
- component models, 439, 443–444
 - transcription, 438, 442–443
 - workflow and use, 439–440, 444
- Embargoes (on data sets), 18, 96, 110. *See also* Language archives
- Encoding of data, 73–74, 77–78, 131, 133, 560
- for repositories, 111
 - Text Encoding Initiative (TEI) guidelines, 322
- Endangered language data
- archiving of, 129, 136, 137–138
 - child language acquisition data (Cree), 392, 393–395, 397–398
 - Electronic Metastructure for Endangered Languages Data, 137
 - untrained forced alignment, 429–432
- Endangered language documentation, 53, 287
- Endangered Language Documentation Program, 136
- lexicography, 310
- Endangered Languages Archive, 136, 467
- Endangered languages movement, 53
- English Lexicon Project (ELP), 416, 570
- Escaping (escape characters), 76
- Ethical approaches to data management, 50–52, 104–105, 123–124. *See also* Ethical practices
- data sharing decisions, 253–254
- Ethical considerations
- data generation and use, 104–105
 - data sharing and use, 123–124
 - Internet data, 123, 475
- Ethically open access, 56
- Ethical practices. *See also* Ethical considerations
- for Aboriginal Australian languages, 341
 - for comparative databases, 341
 - for corpus data, 216, 464–465
 - Data Guardians (DGs), 198
 - for data recording, 259, 291, 464–465
 - for data sharing, 20, 56, 96–97, 125–126, 198–199, 341
 - for data use, 11, 217
 - ethical fieldwork model, 52
 - Ethics Statement* (Linguistic Society of America), 52
 - for Indigenous languages, 49, 51–52, 52–54, 54–55, 428
 - for interview data collection, 217
 - for language documentation, 52–54, 57, 271, 428, 432
 - for legacy data, 51–52
 - for marginalized community data, 253, 271
 - for personal data, 104–105, 123, 464–465
 - repatriation of data, digital, 105
 - for sharing ethnographic data, 253–254
 - for signed language data, 269–270, 370, 464–465, 472

- Ethics approval (Canada), 402
- Ethics Statement* (Linguistic Society of America), 52
- Ethnobiological terminology, 306
- Ethnobotanical knowledge, 135
- Ethnographic research, 249
 - closed data practice, 253–254
 - data citation practices, 253
 - data collection, 250
- Ethnography of communication, 40, 249
- Ethnologue, 40
- EU General Data Protection Regulation (GDPR), 94, 96, 198–199, 398
- Evaluation of scholars and research, 160, 178. *See also* Review, promotion, and tenure (RPT) process
 - best practices for evaluators, 167–168, 180
 - best practices for institutions, 168
 - Declaration on Research Assessment (DORA), 90, 158, 172
 - for hiring, 17, 162
 - Leiden Manifesto (on research evaluation), 158, 172
 - metrics-based, 158
 - overreliance on publication metrics, 161
 - for tenure and promotion, 162, 172
- Experimental research
 - data analysis, 543–544, 562
 - data collection, 541–542
 - data management, 535–536, 540, 558, 560
 - data repositories, 540, 544, 563
 - eye-tracking, 30, 544, 561
 - file organizational structure, 540, 542, 558
 - Mechanical Turk (crowdsourcing website), 541, 542
 - perception tasks, 557
 - scripts (computer), 561, 562
 - stimuli, 558–560
 - web-based experiments, 539
- Experimental semantics/pragmatics, 539
- Experimental syntax, 531–533
 - counterbalancing, 534
 - data types, 532, 534
 - demographic data, 535
- Experts, language, 627
- Extralinguistic knowledge, 135, 136
- Eye-tracking, 30, 544

- FAIR data principles (findable, accessible, interoperable, reusable), 94, 144, 345
 - adherence to, 448
 - benefits of adherence to, 447, 451
 - statistical code, 447
- Feature structures (Head-Driven Phrase Structure Grammar), 35–36
- Fieldwork
 - batteries and power banks, 330, 333
 - for child language acquisition study, 393, 394
 - collaborative, 303
 - collection of audio recordings, 291, 328
 - collection of video recordings (signed language data), 269–270
 - corpus phonology, 582
 - daily routine for, 328, 329, 334
 - data practices, 131, 245, 288
 - data storage practices for, 66, 329, 331
 - data workflow for, 62, 64, 106, 288, 327–331, 525
 - elicitation, 288–289, 305, 328, 514, 527–528
 - equipment management, 330
 - ethical practices, 124, 269–270, 271, 432
 - field notebooks, 328, 330, 331
 - field notebooks, digital, 297
 - file naming, 62, 64, 524
 - judgment elicitation, 514
 - metadata, 288, 328
 - signed languages, 269–270, 271
 - solar panels, portable, 330
 - theoretical syntax, 526, 527–528
 - time frame and project scale, 302
 - tools for, 296, 297, 328
 - training of fieldworkers, 238
- File compatibility
 - cross-platform, 77–78
- File drawer problem, 15–17, 118
- File folder structure, 65–66, 107, 108, 460, 540, 558
- File formats
 - accessibility of, 64
 - corpus files, 189
 - CSV files, 75, 415, 449
 - database file formats, 202, 335, 524, 620, 640
 - file conversion, 73, 74, 78
 - file conversion, batch, 78–79, 466
 - file migration, 106
 - HTML, 75, 76, 458, 542
 - human-readable versus machine-ready, 75, 153, 415, 533
 - JavaScript Object Notation (JSON), 533, 620
 - locale settings, 77–78
 - lossless formats, 65, 106, 260
 - Microsoft Excel spreadsheet files, 75–77, 613
 - Microsoft Excel used for coding, 216, 251, 516, 613–615
 - Microsoft Word documents, 75–77, 84
 - natural language processing (NLP), 415
 - persistent formats, 95, 448–449
 - plain text files, 75–78
 - proprietary, 74, 75, 106
 - proprietary versus open, 64, 75, 340
 - recommended file formats, 64–65, 75–77, 448–449
 - for repositories, 64–65, 111, 448–449, 620, 640
 - RIS (Research Information Systems) format, 152
 - software for documentary fieldwork, 296, 328
 - spreadsheets, 75, 331, 333, 458, 613
 - sustainable formats, 64–65
 - tab-delimited, 65, 74, 75–77, 449, 456
 - text files, character and line encoding of, 73–74, 77–78, 558
 - transferring and overwriting files, 329
- FileMaker Pro, 335, 339–340, 635

- File migration, 106. *See also* File formats: file conversion
- File naming, 108–109, 281
- corpus files, 189, 407
 - corpus linguistics, 460
 - data sharing and reusability, 449
 - derivative files associated with recordings, 251
 - ELAN annotation files, 378–379
 - fieldwork, 62, 64, 524
 - and folder structure, 65–66, 74, 78–79, 260, 460, 542
 - for language archives, 288
 - readme files, 449
 - repository schemas for, 108
 - semantic file naming, 524
 - templates, 441
 - and workflow, 64, 216, 251, 407
- Findability of data, 144, 163, 164–165. *See also* FAIR data principles
- Cross-Linguistic Data Formats (CLDF) initiative, 351
 - metadata, 448
- First language acquisition. *See* Child language acquisition data
- First Nations language data. *See* Indigenous language data
- First Nations Principles of OCAP (Ownership, Control, Access, Possession), 55
- FLEx (FieldWorks Language Explorer), 287, 293–296, 332
- lexical database records, 307–309, 311–312
 - pros and cons for lexical database, 306, 321
- fMRI (functional magnetic resonance imaging), 30, 547–548, 549
- neuroimaging meta-analysis, 550, 551
- Forced alignment, 225–226, 423–424
- manual checking, 227, 431
 - tools and methods for, 424
 - tools for, 195–196, 243, 431
 - untrained, 429–432 (*see also* Untrained forced alignment)
- Funding requirements
- for data availability, 151
 - for Data Management Plans (DMPs), 102, 113
 - for data sharing and ownership, 122
- General Data Protection Regulation (GDPR), 94, 96, 198–199, 398
- Generative approaches to linguistics, 524, 527, 528
- Geocoding, 241, 335
- both languages and varieties, 342
- Git (version control system), 84, 536, 540
- Git Bash (command-line emulation layer), 83
- GitHub (repository hosting service), 84, 540
- Git Large File Storage (Git-LFS), 544
 - size limits, 544
- Global Englishes, cross-corpus analysis of, 206
- Global Signbank, 272, 467
- Glossing, 31, 32
- Leipzig Glossing Rules, The, 31, 619
 - signed language data, 272, 367–368
 - theoretical syntax, 525
- Glottolog, 40
- Glottocodes, 342, 347, 590, 632
- GOLD ontology (General Ontology for Linguistic Description), 137
- Grammaticality judgments, 29, 41, 514, 518. *See also* Elicitation
- Grammatical relations, 634–635
- HARKing (hypothesizing after results are known), 17
- Hawai'i Creole (Pidgin), 222
- Hawai'i language history, 221–222
- Historical linguistics, 327, 333. *See also* Comparative research;
- Language change
 - Bayesian phylogenetic methods, 349–351
 - cognate forms, 333, 338, 339, 347
 - cognate sets, 345, 348
 - comparative databases, 335, 340
 - computational phylogenetics, 345
 - concept sets, 347
 - Cross-Linguistic Data Formats (CLDF) initiative, 347–348
 - historical databases, 335–338
 - typological reference languages, 339
 - word lists, uses of, 339
- HTML, 75, 76, 458, 542
- Human subjects. *See* Institutional review boards
- Iconicity norms, 492, 492–493
- Identifiers
- for data accessibility, 448
 - digital object identifier (DOI), 94, 147, 448
 - persistent identifiers (PIDs), 94, 147, 150, 153, 448, 476
 - recommendations for, 95–96, 147
 - used in citations, 147, 148, 149, 448, 615
 - version numbers, 147
- Identity
- of participants, 40, 41
 - of signers, 474, 475
 - of speakers, 222
 - scholar digital identity, 161
- Impact of research. *See* Research impact
- Indigenous data sovereignty (IDS), 54–56, 315
- CARE Principles for Indigenous Data Governance, 55–56
 - First Nations Principles of OCAP (Ownership, Control, Access, Possession), 55
 - Indigenous Data Sovereignty (IDS) movement, 54
 - and open access, 56, 432
- Indigenous language data, 49
- anonymization and deidentification of, 398
 - archival documentation, 317, 318, 322
 - Chisasibi Child Language Acquisition Study (CCLAS), 392–393
 - collected for theoretical syntax, 523
 - data sharing and access, 428–429
 - “data” used as term for, 315–316
 - ethical practices for, 51–56, 428–429
 - first language acquisition data, 392, 399

- Indigenous data sovereignty (IDS), 54–56, 315, 432
- Indigenous Language Digital Archive, 322
- Indigenous Tweets project, 481, 482–484, 486
- intellectual property, 120, 124, 428
 - ownership of, 50–52
 - protocols for use of, 107, 112
 - untrained forced alignment, 423–424, 427–429, 429–432
- Indigenous languages, 49
 - language reclamation, 315–316, 323, 324
 - language revitalization, 135, 315–316, 481
- Indigenous research methods, 54
- INESS (Infrastructure for the Exploration of Syntax and Semantics) (treebank tool), 501–502, 505–507
 - citation, 507–510
 - metadata, 503–504, 505–506
- Information density (of language data), 74
- Information processing, three levels of detail, 358
- Informed consent, 104–105
 - CA recordings, 259
 - corpus data, 187–188, 217, 269–270, 402–403, 464–465
 - of parent or guardian, 402–403
 - reconsenting protocols, 370
 - recorded orally, 269, 328
 - regarding sharing of research data, 123, 254
 - signed language video data, 269–270, 370, 472, 464–465
- Inherent intellectual property (IP), 104. *See also* Intellectual property (IP)
- Inheritance plans (for data sets), 110
- Institutional review boards, 104. *See also* Informed consent and data sharing, 126
 - Ethics approval (Canada), 402
 - Internet data, 475, 476
 - signed language data, 472, 475
- Instrumental data, 30
- Intellectual property (IP), 104. *See also* Copyright law
 - Indigenous language data, 120, 124, 428
 - patents, 118
 - sui generis* rights, 119
 - trademarks, 118
- Interannotator agreement, 83, 501, 606
- Interdisciplinary use of linguistic data, 42
- Interlinear glossing, 32
 - as parallel data, 131
 - tools for, 293–295, 296
- Internet data
 - citation of, 476
 - compared to open data, 476
 - corpora, 485
 - ethical considerations, 475
 - genre, 475, 477
 - Indigenous Tweets project, 481, 482–484, 486
 - metadata, 475, 484–485
 - permission, 475, 476
 - privacy, 475, 485
 - reproducibility, 476, 485–486
 - sampling, 473–475
 - signed language data, 273, 473, 475, 476–478
 - Terms of Service, 485–486
 - Twitter, 481–483, 484–486
- Interoperability of data, 125, 144, 347, 448, 621. *See also* FAIR data principles
 - Scholix initiative (to facilitate the exchange of data citations), 166
- Introspective data, 20, 494, 618. *See also* Elicitation; Judgments
- Intuition, data based on, 492–495
- Iquito Language Documentation Project, 307
- IRBs (Institutional review boards), 104. *See also* Informed consent and data sharing, 126
 - Ethics approval (Canada), 402
 - Internet data, 475, 476
 - signed language data, 472, 475
- ISCAN (Integrated Speech Corpus Analysis), 195, 196, 200–206
- Ishi (speaker of Yahi), 135–136
- ISLE Meta Data Initiative (IMDI) standard, 42, 136
- ISO 639 language codes
 - and data integrity checking, 592
 - and other common codes, 342, 632
 - used in file naming, 524
- Joint Declaration of Data Citation Principles (FORCE11), 143–144
- Journal impact factor (JIF), 157, 160, 161, 172
- Judgments. *See also* Elicitation
 - acceptability, 29, 517, 531
 - acceptability experiments, 531–533, 534
 - elicitation of, 514
 - formal linguistics, 618
 - grammaticality, 29, 41, 518
 - introspection, 20, 494, 618
 - judgment-based research, 29–30, 518
 - levels of confidence, 625, 626
 - metadata, 519
 - norming of subjective judgments, 490, 492, 493–494
 - speech perception, 567
- Julia (programming language), 360–364
- Kaipuleohone Language Archive, 134, 288, 297
- LaBB-CAT (Language, Brain and Behaviour—Corpus Analysis Tool), 225–226
- Labov, William, methods of, 13, 210
- Language* (journal), 4
- Language acquisition. *See also* Child language acquisition data; Second language (L2) acquisition
 - CHAT transcription (Codes for the Human Analysis of Transcripts), 370
 - child language acquisition, 369, 370–371, 385–386, 392
 - child language corpora, 385–389, 392–393
 - computational models of, 357
 - signed language acquisition, 369, 371–373

- Language and literacy resource materials
 - accessibility of, 283, 341
 - community-oriented, 429
 - examples of, 279–280, 301
 - for language reclamation, 323
 - sharing draft materials, 303
- Language Archive, The (TLA), 136, 467
- Language archives, 51, 53. *See also* Data repositories
 - Archive of the Indigenous Languages of Latin America (AILLA), 92, 96, 134
 - citation of, 149–150
 - digital archives, 15–16, 137
 - Digital Endangered Languages and Musics Archives Network (DELAMAN), 137
 - Endangered Languages Archive, 136, 467
 - funding of, 137, 138
 - Indigenous Language Digital Archive, 322
 - Kaipuleohone Language Archive, 134, 288, 297
 - Language Archive, The (TLA), 136, 467
 - metadata, 288–290
 - PARADISEC (Pacific and Regional Archive for Digital Sources in Endangered Cultures), 92, 134, 331
 - regional archives, 134, 137
 - signed language data sets, 268, 467
 - use by future linguists, 134, 138
- Language as a social construct, 598
- Language catalogs, 40
 - Ethnologue, 40
 - Glottolog, 40, 347, 590
- Language change. *See also* Historical linguistics; Language variation
 - computational models of, 355–357, 359–360
 - signed language change, 463, 477
- Language description. *See* Language documentation
- Language documentation
 - archival documentation, 315–317, 318, 322
 - archives for, 134, 136 (*see also* Language archives)
 - collection of audio recordings, 280–281, 289, 291
 - community data access, 270, 428
 - corpus transcription, 427–429
 - creation of language and literacy resource materials, 279, 283, 302, 323, 429
 - creation of language dictionaries, 295, 301, 302, 310
 - data management practices, 277, 278, 282, 283, 291
 - data of, 28, 289
 - endangered languages, 136, 287 (*see also* Endangered language data)
 - equipment management, 330
 - ethics of, 52–54, 57, 271, 428–429, 432 (*see also* Ethical practices)
 - fieldwork, 288, 327–331
 - file naming, 281
 - GOLD ontology (General Ontology for Linguistic Description), 137
 - Indigenous languages, 49, 52–54, 57, 111–112, 428–429, 432 (*see also* Indigenous language data)
 - language work, 278
 - lexical data, 304, 310
 - local and distant documentation, 267–268
 - metadata, 279, 288
 - sharing of language and literacy resource materials, 283, 301, 429
 - sign languages, 267–268, 271, 468
 - tools for, 288, 296, 297
 - workflow, 282, 288, 331
- Language endangerment, 49, 53. *See also* Endangered language data
- Language experts, 627
- Language names, standardization of, 342
- Language reclamation, 52, 315–316, 322, 324
 - goals of, 315, 323
 - reclamation-driven reconstruction, 322
 - reconstitution, 322
- Language revitalization, 137–138, 428
 - archives, 137–138, 283
 - language documentation used for, 135, 278
 - literacy materials, 279–280
 - and reclamation, 315
- Language variation, 215–216, 283
 - Hawai'i Creole (Pidgin), 221
 - regional, 238
 - signed languages, 477
 - theoretical linguistic fieldwork, 528
 - Uzbek, 513–514, 517
- LaTeX
 - example formatting, Python script for, 525
 - pros and cons, 307–308
- Legacy data, 138
 - collected into a corpus, 186
 - encoding of, 322
 - ethics of, 51–52
 - recording quality, 227
- Leiden Manifesto (on research evaluation), 158, 172
- Leipzig Glossing Rules, The, 31, 619
- Lexical data, 36–38, 301
 - English Lexicon Project (ELP), 416, 570
 - historical comparative database, 335–338
 - input into FLEx, 332
 - for language and literacy resource materials, 302
 - lexicon structures, 37
 - representations of, 37–38
 - signed language database, 375–377, 467
 - types of, 304
 - validity of, 310
- Lexicography, 37, 304
 - data of, 304
 - and language documentation, 301
 - methodology and workflow, 304, 307–309, 310, 311–312
 - objectives, 302

- outputs of, 301, 306
 - tools for, 306
- Lexicons, structures of, 37
- Lexicostatistics, 348
- Licensing of data, 94, 95, 111, 124. *See also* Copyright law
- Linguistic annotation. *See* Annotation
- Linguistic data. *See also* Annotation; Corpora; Data
 - acceptability judgments, 29, 531–533
 - definition of, 49–50
 - documentary linguistic data, 28, 288
 - ethics, 50 (*see also* Ethical practices)
 - example format and typesetting, 525–526
 - format for deriving presentation forms, 525
 - Indigenous approaches to, 50 (*see also* Indigenous language data)
 - interdisciplinary use of, 42
 - language catalogs, 40
 - lexical data, 36–38, 304
 - metadata, 40–42 (*see also* Metadata)
 - socially stigmatized variety, 217
 - structural analyses, 33–36, 50
 - transparency of, 5, 10–11
- Linguistic description. *See* Language documentation
- Linguistic examples
 - citation of data source, 31, 262, 297
 - formatting and typesetting of, 525–526
- Linguistic metadata, 40–42. *See also* Metadata
- Linguistic norms, 533
 - concreteness, 489, 491
 - construct validity, 493
 - emotional valence, 489, 490, 491, 492
 - figurative language, 492
 - iconicity, 492–493
 - norming data sets, 490, 491–492
 - reproducibility, 492–494
 - semantic norms, 489, 490, 491–492
 - sensory and perceptual norms, 491
 - word association norms, 549, 550
- Linguistic research. *See* Research
- Linguistics, generative approaches to, 524, 527, 528
- Linguistics as a social science, 10–11, 357
- Linguistics Data Interest Group (part of the Research Data Alliance), 5, 144
- Linguistic Society of America *Ethics Statement*, 52
- Linguistic typology. *See* Typology
- Locators. *See also* Identifiers
 - used in citations, 148, 149
- LOCKSS (Lots of Copies Keeps Stuff Safe), 106, 130, 281, 341
- Longitudinal research
 - child language acquisition, 385–389, 392–393
 - L2 learner corpora, 412
 - sociophonetic, 221
- Long Now Foundation, 130
- Long-term archiving, 129, 130
- Long-term preservation, 41, 130–133, 137–138
 - file formats for, 65
 - selection of data for, 63
- Lyon Corpus (child language acquisition, French), 386, 388
- Machine learning, 196, 468, 501, 548, 549
- Machine readability, 31, 33, 108
 - file formats, 75, 153, 413–414
 - signed language data, 268, 272, 367, 374
- Make Data Count project, 173, 179
- Marr, David, levels of information processing, 358, 360
- Mechanical Turk (crowdsourcing website), 541–543
- Metadata, 40–42
 - aggregation of, 168, 589
 - annotation of, 32
 - attribution of contributors in, 146
 - child language acquisition, 374, 379
 - citation of elements missing from, 146
 - collaborative (team) projects, 303
 - Component Metadata Infrastructure (CMDI), 503
 - corpus metadata, 189, 407, 464
 - for data citation, 145, 153
 - data dictionaries, 68, 108
 - Data Documentation Initiative standard, 67
 - DataCite schema, 67, 146
 - definition of, 41, 67
 - demographic data, 189, 211, 223, 279, 615–616
 - documentary linguistic metadata, 279, 281, 288–289, 328
 - documentation of, 66–68, 107–109, 558, 561
 - Dublin Core schema, 67, 109
 - encoded in file and directory structure, 79
 - experimental research, 533, 534, 559, 560
 - fieldwork data, 288–289, 328, 331
 - file naming, 65–66, 108–109, 251, 281, 449
 - findability, 448
 - good practices for, 64, 153, 297–298, 558
 - harvestable, 94, 448, 503
 - historical database metadata, 338
 - importance of, 42, 95, 289, 593
 - inclusion of references to publications in, 95, 149, 164
 - Internet data, 475, 484–485
 - ISLE Meta Data Initiative (IMDI) standard, 42, 136
 - judgment-based research, 519
 - keywords, 448
 - machine-readable, 108, 153, 503
 - Metadata Object Description Schema, 109
 - Microsoft Excel used for, 251–253, 407 (*see also* Spreadsheets: metadata)
 - Open Language Archives Community (OLAC) standard, 42, 67, 136–137, 592
 - readme files, 67–68, 108, 449, 558
 - repository standards for, 67–68, 94, 153, 448
 - schemas for, 67–68, 109, 136, 281
 - signed language data, 464–465, 467, 475
 - software version numbers, 298

- Metadata (cont.)
 spoken (on recording), 291
 standards for, 42, 67–68, 94, 464, 467
 treebanks, 503
 types of, 107–109
 video data metadata, 251–253, 260, 331, 374, 378–379
 workflow for, 108–109, 329, 331, 615–616
- Meta-research, 4
- Metrics Toolkit, 158
- Modeling. *See* Computational modeling
- Morphological analysis
 annotation of, 31, 332, 387, 394–395, 612–613
 parsing in FLE_x, 332
- Morphosemantic analysis, 609, 612, 613
- Multilingual communities, 392, 398, 430
- Multilingual data, 346–348, 392, 430, 611, 612, 613
- National Digital Stewardship Alliance (NDSA), 66
- Native speaker or signer experts, 620
 data reliability, 625
- Naturalistic data, 28
 child language acquisition, 371–372, 392, 398
 signed language data, 273, 371–372, 465, 472
- Natural language processing (NLP)
 computational modeling, 356
 for corpora, 413–414
 file formats, 415
 and L2 learner data, 411–412, 414, 418
 NLP tools, 82, 414, 415, 417–419
 Suite of Automatic Linguistic Analysis Tools (SALAT), 414, 419
 text length, 414
 Tool for the Automatic Analysis of Lexical Sophistication (TAALES), 415, 418–419
 untrained forced alignment, 423–424, 427–429, 429–432
 uses of, 138, 411–412, 548
- Natural Language Toolkit, 83
- Naturally occurring data, 257
- Neurolinguistics. *See* Computational neurolinguistics
- Neurosynth (platform for fMRI meta-analysis), 550
- Nonbinary speakers, sociophonetic data from, 229
- Norms. *See* Linguistic norms
- NVivo (qualitative data analysis software), 611, 612–613
- Observer's paradox, 28, 210, 250, 269, 472
- OCR (optical character recognition), 321
- Online collection page for data management use cases, 7
- Online open access companion course, 6–7
- Online Speech/Corpora Archive and Analysis Resource (OSCAAR) (now SpeechBox), 241
- Open access corpora, 185–186, 189, 190, 197–198, 468
- Open access data sets. *See also* Data sets
 curriculum vitae, inclusion on, 16
 fMRI data, 547–548, 552
 licensing and permission, 124–126
 publicizing availability of, 566
 speech perception data, 565–566
 typology, 639–640
- Open access movement, 11, 53
 ethically open access, 56
- Open data. *See also* Open access data sets
 Cross-Linguistic Linked Data (CLLD) project, 351
 Internet data compared to, 476
- Open data movement, 53. *See also* Open access movement
 ethical limitations of, 253, 254
 exceptions to, 56
- Open data practice, 11, 125
 accessibility of data, 9, 151, 476
 availability of metadata, 153
 benefits of openness, 19, 162, 341, 447, 518
 best practices for evaluation of scholarship, 17, 167–168
 best practices for journals and publishers, 168
 best practices for scholars, 97, 167, 544
 ethical considerations for, 56, 96–97, 253
 Peer Reviewer's Openness Initiative, 21
 repository access control, 96–97
 restricted data, 336
 shift to culture of, 14, 20–21, 173
 signed language video data, 380
- Open Language Archives Community (OLAC)
 metadata standard, 42, 67, 136–137, 592
 OLAC Role Vocabulary, 146
 repositories, 93
- Openness. *See also* Open data practice
 benefits of, 19, 162, 341, 447, 518
 culture of, 14, 20–21, 173
 repository access control, 96–97
- OpenRefine (data cleaning and transformation tool), 611, 613
 used with Microsoft Excel, 611, 613, 614–615
- Open Science Framework (OSF), 16, 536, 544
- Open source code, 432, 447, 450–451. *See also* Scripts
- Operationalization, 598
- ORCID (Open Researcher and Contributor ID), 161
- Orthography
 for ASL (American Sign Language), 368
 for archiving data, 519
 to distinguish homographs, 213
 used in transcription, 213, 319–321
 writing systems, 132–133
- OSF (Open Science Framework), 16, 536, 544
- Ownership
 community ownership of linguistic data, 50–52
 of data, 20, 104, 123, 475
 First Nations Principles of OCAP (Ownership, Control, Access, Possession), 55
- Palatograms, 30
- Paradigmatic analyses, 34–35, 36
- PARADISEC (Pacific and Regional Archive for Digital Sources in Endangered Cultures), 92, 134, 331

- Parallel data, 131–132
 corpora, 386
 treebanks, 499
- Participant identity. *See* Identity
- Participant observation, 249
 for documentary linguistics, 306
- Paywalls, 95, 379
- Pear Film, The, 132
- Peer review
 CA (Conversation Analysis) data sessions, 261
 of data sets, 16, 96
 Peer Reviewer's Openness Initiative, 21
 peer-reviewed journal articles, 172
 preparation of papers for, 16, 149
 preprints, 20, 172
 of registered reports, 18
- Permissions. *See also* Copyright law
 for data sharing, 125, 198
 for data use, 121, 124
- Persistent identifiers (PIDs), 94, 448. *See also* Identifiers
 Handle system, 153
 International Standard Language Resource Number, 153
 for Internet data, 476
 for location within a corpus or archive, 150
 recommendations for, 147
 for treebank data, 507
- Personal data. *See also* Consent
 archiving of data including, 96–97
 consent, 269–270, 328
 consent and anonymization, 104–105, 189, 259
 Data Guardians (DGs), 198
 ethical considerations for, 123, 217, 259, 269–270, 464–465, 475
 processing and storage of, 251–253, 278
- p-hacking, definition of, 17
- PHOIBLE (Phonetics Information Base and Lexicon Database), 589–591, 592
 data sharing, 592
 PHOIBLE Online, 592–593
- PhonBank (child phonology database), 388, 391, 392, 398
 Phon (open-source software), 391, 395–397
- Phonetic data analysis, 195–197, 203–205
 database “enrichment”, 202–203
- Phonetic identification data, 569
- Phonological CorpusTools (PCT), 575
 analysis algorithms, 578–581, 585
 original/fieldwork data, 582–585
 preexisting corpora, 576
 replicability, 575
- Phonological data
 analysis of, 395–397, 579, 585
 corpus phonology, 575, 576, 582–585
 Irvine Phonotactic Online Dictionary of English (IPhOD), 576
 perception experiments, 557, 558–560
 PhonBank (child phonology database), 388, 391, 392, 398
 pronunciation variants, 584
 TextGrids, 396, 584
 transcription of, 394
- Phonological inventory data
 PHOIBLE (Phonetics Information Base and Lexicon Database), 589–591, 592
 phonological language descriptions, 590
 UPSID (UCLA Phonological Segment Inventory Database), 589
- Plain text files, 75–78. *See also* File formats
- Polyglot (hybrid database format), 202, 203
- Postprocessing of data, 81–82, 85
 tokenization, 82
 use of Natural Language Processing (NLP) tools for, 82
- Praat (speech analysis software)
 file formats, 65
 Praat TextGrid files, 226, 391, 426, 584
 scripts for automatic measurements, 196, 228
- Preregistration (of methods, hypotheses), 18, 494, 534, 536, 544–545
- Primary data. *See also* Data
 definition of, 50
 signed language data, 476
 stratum in database, 336
 text corpora as, 29
- Priming effects, controlling for, 455, 457
- Privacy, 51, 80, 104. *See also* Confidentiality; Data security
 archiving data, 96–97
 corpus data, 217
 Internet data, 123, 485
 Internet videos, 475
 sign language video data, 270, 370, 464–465, 468, 472
- Processed data. *See also* Data
 archiving in a repository, 96
 citation of, 151
 publication of, 151
- Programming languages
 code, documentation of, 361–363, 447–451, 457–458, 535, 545, 561–562
 code, open source, 432, 447, 450–451
 code, sharing of, 447, 494, 532, 535, 585, 639–640
 for corpus linguistics, 454
 Julia, 360–364
 Markdown (markup language), 534
 practices for code clarity, 360–364, 450–451, 457–458, 535, 562
 practices for coding, 456, 535, 545, 561–562
 pseudo-code, 362–363
 Python, 83, 282, 307–308, 360
 Python script for LaTeX example formatting, 525
 R (software environment), 83, 282, 535 (*see also* R (software environment))
 R (software environment), packages for, 543–544
 R Markdown (format for R code documentation), 535, 545
 regular expressions (regex), 80, 454, 526, 613

- Promotion process. *See* Review, promotion, and tenure (RPT) process
- Providence Corpus (child language acquisition, American English), 386, 387–388
- Psycholinguistics
 - acceptability judgment experiments, 531–532
 - archiving R scripts, 535, 545
 - data management, 489, 532, 540–542, 543–544, 557
 - experimental data, 532, 541–542
 - experimental scripts, 561
 - file formats, 535–536
 - lexical decision tasks, 557, 570
 - participant metadata, 560
 - semantic norms, 489, 491–492, 493, 533
 - stimuli, 558–560
 - web-based experiments, 539, 541–542
- Psychology. *See also* Psycholinguistics
 - social psychology, 9, 14, 21
- Publications
 - bias against negative results in, 15
 - data accessibility, 90, 109
 - data sets as, 19, 109, 173, 178
 - linking to data, 164, 166, 167
 - preference for novelty in, 15
 - preprints, 16, 172–173
 - replications, rates of, 12
 - representativeness of, 15
 - signed language research, 368, 476
 - web publications listed on CV, 627
 - with open access data sets, 19, 90, 109, 566, 571, 638
- Public corpora, 185–186, 189, 197–198, 468. *See also* Open access data sets
 - sharing of, 190, 198, 468
- Python, 83, 282
 - pros and cons, 307–308, 360
 - script for LaTeX example formatting, 525
- R (software environment), 83, 282, 535
 - annotation of code, 450–451, 457, 562
 - for corpus linguistics, 454–460
 - R Markdown (markup format), 535, 543, 545
 - R packages, 543–544
 - RStudio (integrated development environment), 543
 - script-writing, 456
 - statistical plots, 460
- Raw data
 - archiving in a repository, 96
 - definition of, 50
 - raw text, 82
 - versus working data, 106
- Readme files
 - Markdown, 534, 536
 - recommendations for, 67–68, 108, 449, 558
- Reciprocity (Indigenous research methods), 54
- Reclamation, language, 52, 315–316, 322, 324
 - goals of, 315, 323
 - reclamation-driven reconstruction, 322
 - reconstitution, 322
- Recordings. *See* Audio recordings; Video data
- Redaction
 - of documentary linguistic recordings, 291
 - of naturally occurring corpus data, 189
 - of naturally occurring data, 74
- Reference management tools, 152
 - BibLaTeX, 152
 - BibTeX, 152, 590, 635–636
 - Research Information Systems (RIS) file format, 152
 - Zotero, 152
- Regional data, 237
 - analysis of, 238
 - collection of, 238, 244
- Reliability, 387, 388, 598, 606
 - interrater reliability, 388, 493, 598, 606
 - native speaker or signer expert data, 625
- Repatriation of data, digital, 105
- Replicability, 11–12. *See also* Reproducibility
 - applied to linguistics, 13
 - data produced for, 41
 - as goal of database, 339
 - introspective methods, 618
 - Phonological CorpusTools (PCT), 575
- Replicable research
 - barriers in linguistics, 447
 - cognitive linguistics, 494
 - corpus phonology, 575
 - methods of, 41
 - semantics research, 494
 - versus reproducible research, 12, 494, 598
- Replication crisis, 14
- Replication data packages, 151
- Replications, conceptual, 12
- Replications, rates of, 12
- Repositories. *See* Data repositories
- Representativeness of data, 11–12, 28
 - corpora speaker sampling, 210
 - corpora speaker subsamples, 212
- Reproducibility, 11–12, 12–14
 - applied to linguistics, 13, 339
 - cognitive linguistics, 491–494
 - computational modeling, 359, 364
 - corpus linguistics, 490
 - data citation for, 143
 - data produced for, 41
 - of database decisions, 338
 - as goal of data sharing, 12, 118, 179
 - inference reproducibility, 14
 - Internet data, 476, 485–486
 - semantics research, 490, 491, 492–494
 - sign language research, 476, 477–478

- Reproducible research, 12–14. *See also* Replicable research
 definition of, 3, 11
 position statement on, 10
 practices for, 18, 339, 361–362, 477–478, 492–494, 540
 transparency of methods, 12, 41, 531, 598
 versus replicable research, 12, 494, 598
- Research
 collaborative approaches to, 53, 198, 303
 community-based, 52–54, 392, 397–399
 dissemination of, 104, 159, 161–162, 397–399
 ethics of (*see* Ethical practices)
 evaluating impact of, 158 (*see also* Research impact)
 meta-research, 4
 outputs of, 4, 16, 19, 172–173, 175, 397–398
 preregistration, 18, 494, 534, 536, 544–545
 reciprocity (Indigenous research methods), 54
 reproducibility of, 3, 12–14 (*see also* Reproducible research)
 sign language communities, practices for, 270–271, 370, 464–465 (*see also* Sign language data)
 validity, 598, 606
- Research assistants
 corpus data collection, 406
 training of, 252–253, 403–404
- Research data. *See* Data; Data sets; Linguistic data
- Research Data Alliance, 5, 125, 144
 Linguistics Data Interest Group, 5, 144
- Research Data Management (RDM), 101. *See also* Data management
 training and courses on, 91
- Research fraud, 447
- Research impact, 157
 altmetrics, 158, 159–160, 162, 166
 bibliometrics, 159
 citation-based metrics, 159, 160
 Declaration on Research Assessment (DORA), 90, 158, 172
 evaluation metrics concepts, 158
 evidence of, 162
 field normalized metrics, 162, 166, 167
 Journal impact factor (JIF), 157, 160, 161, 172
 Leiden Manifesto (on research evaluation), 158, 172
 limitations of metrics, 158, 160, 163, 167
 publication metrics, 159, 160, 162
 validity of metrics, 160
- Reusability of data, 144, 449. *See also* FAIR data principles
 licensing, 94, 95, 111, 124–126
- Reuse of data, 61. *See also* Reproducible research
 archive restrictions on, 336
 barriers to, 64
 legal considerations, 104, 124–126
 value of, 69, 447, 634
- Review, promotion, and tenure (RPT) process, 171–172.
See also Evaluation of scholars and research
 data-related outputs, 109, 177–178
 practices for evaluators, 162, 180
 practices for scholars, 162, 179
 scholarly outputs, 172–173, 175, 178
- Revitalization, language, 137–138, 428
 archives, 137–138, 283
 language documentation used for, 135, 278
 literacy materials, 279–280
 and reclamation, 315 (*see also* Language reclamation)
- Rosetta Stone, 130–131
- Santa Barbara Corpus of Spoken American English, 197, 198, 257
- Sapir, Edward, 135–136
- SayMore (language documentation tool), 287, 293, 296, 328
 transcripts imported into FLEx, 332
- Scalability of research, 490
- Scholar digital identity (scholarly profile), 161
- Scholarly outputs, 172–173, 175
 data management and sharing, 161–162, 173, 177, 397–399
 database contributions, 627
 nontraditional, 172, 178, 254
 recognized for review, promotion, and tenure, 109, 172–173, 176–177, 178
- Scholarly profile (digital identity), 161
- Scholix initiative (to facilitate the exchange of data citations), 166
- Science Citation Index, 157
- Scientometrics, 157
- Scottish Corpus of Texts and Speech (SCOTS), 197–198
- Scripts (computer). *See also* Programming languages
 documentation of, 361–363, 447–451, 457–458, 535, 545, 561–562
 open source code, 432, 447, 450–451
 practices for coding, 456, 535, 545, 561–562
 pseudo-code, 362–363
 Python script for LaTeX example formatting, 525
 sharing of, 447, 494, 532, 535, 585, 639–640
- Second language (L2) acquisition
 combined corpora, 413
 data sharing, 408
 English as a Second Language (ESL) corpora, 401, 402, 405–406
 IRIS digital repository, 405, 408
 L1 influence, 412
 L1 influence, language distance measure, 416
 learner corpora, 411, 412–414
 natural language processing (NLP), 411–412, 415–416, 418
 proficiency measures, 404, 405
 speech perception, 569
 statistical analysis, 415–416, 417
 Suite of Automatic Linguistic Analysis Tools (SALAT), 414
 word recognition (naming) task, 416
- Security of data, 103, 105–107, 110–112
 General Data Protection Regulation (GDPR), 94, 96, 198–199, 398

- Security of data (cont.)
 personal information, 96–97, 105, 123, 259 (*see also* Personal data)
 repository access control, 96
 video data, 252–253, 380, 472
- Semantics
 associative concept dictionaries, 549
 construct validity, 493
 data sharing, 494
 lexiconeural mapping, 548
 neurosemantic modeling, 547–548, 549–550
 polysemy, 493–494
 replicability, 494
 reproducibility, 490, 491, 492–494
 semantic norms, 489, 490, 491–492 (*see also* Linguistic norms)
 semantic prosody, 489–490, 491
 sensory and perceptual language, 490
- Semantic tiling, 548
- Sharing. *See* Data sharing
- Sign language avatar technology, 270
- Sign language data. *See also* ASL (American Sign Language)
 annotation of, 374, 465–467
 anonymization of, 270
 archiving, 467, 476
 ASL Signbank, 369, 375–377
 ASL-LEX database, 376
 Auslan Signbank (Australian Sign Language), 467
 bimodal bilingual acquisition, 369
 Binational Bilingual Bimodal (BiBiBi) project, 369
 BISINDO corpus (Indonesian Sign Language), 268, 272
 Black ASL Project, 474
 British Sign Language (BSL) Signbank, 376
 child language acquisition, 369, 371–373
 citation of, 476
 corpora, 268, 463–468
 Corpus NGT (Sign Language of the Netherlands), 465–466, 468
 ELAN used for, 374, 378–379, 465–467
 elicited, 472, 473
 ethical considerations, 269–270, 370, 464–465, 472, 474
 Global Signbank, 272, 467
 glosses for, 272, 367–368, 471
 iconicity norms, 492
 ID glosses, 369, 374, 376, 377–378, 466
 iLex (annotation software), 465, 466
 informed consent, 269–270, 369, 464–465, 472, 476
 Internet data, 273, 473, 475, 476–478
 language change, 463, 476
 language contact, 273, 467
 machine readability, 268, 272, 367, 374
 metadata, 272, 374, 475
 metadata standards for, 464, 467
 metalanguage for glossing, 272
 naturalistic, 273, 372, 465, 472
 nonmanuals, 466
 reproducibility, 476, 477–478
 sharing of, 273, 379–380, 467, 476
 signbank databases for, 375, 467
 underdocumented varieties, 474
 video data collection, 269–270, 370–371, 372, 464–465, 472
 written representations of, 272, 367–368, 471
- SLAAASh (Sign Language Acquisition, Annotation, Archiving and Sharing), 369, 374
- Slack (collaboration tool), 540
- Social media
 corpora, 485–486
 ethical considerations, 475
 Indigenous Tweets project, 481, 482–484, 486
 metadata, 475, 484–485
 sign language data, 475
 translations, 486
 Twitter, 481–484, 485–486
 used for data sharing, 379
- Social networks, computational modeling based on, 355–356
- Social psychology, 9, 14, 21
- Social sciences
 computational modeling in, 357
 definition of, 10
 experiment design, 358
 lack of transparent research in, 9–10
 replication in, 12
 scientific method practiced in, 356
- Sociocultural linguistic research, 249
- Sociolinguistic Archive and Analysis Project (SLAAP), 186, 241
- Sociolinguistic data, 40
 analysis of, 215–216, 281–282
 corpora, 185–186, 209–210
 corpora metadata, 189–190, 211–212
 ethical considerations for, 189, 216–217
 Indigenous languages, 486–487
 legacy materials, 186, 227
 linguistic variation, 215–216, 238, 280, 283, 477
 questionnaires used for, 278–279
 sign languages, 477–478
 sociolinguistic interviews, 210–211, 223–225
- Sociolinguistics Lab at the University of Ottawa, 209–210, 217–218
- Sociolinguistics Server (SOLIS) at the University of Hawai‘i at Mānoa, 225, 230
- Sociophonetic data
 collection of regional data, 238–239, 244
 data processing, 224–229, 242–243
 reporting data, tools for, 229–230
 speaker identity, 222
 storage of, 230
 used to study variation and change, 221–223, 238
 vowel perception tasks, 238, 239–240
- Software. *See also* R (software environment)
 computational phylogenetics, 347
 concordance software, 215

- Dropbox, 378, 379
- ELAN (annotation tool), 32–33, 65, 282–283, 296, 378, 465–467 (*see also* ELAN)
- FileMaker Pro, 335, 339–340, 635
- FLEx (FieldWorks Language Explorer), 293–295, 306, 307–309, 311–312, 321, 332
- Git (version control system), 84, 536, 540
- Indigenous Language Digital Archive, 322
- LaTeX, 307–308, 525–526
- LibreOffice Calc, 458, 459
- machine learning–based phonetic software, 196
- Microsoft Excel, 81, 407, 459, 516, 613
- NVivo (qualitative data analysis software), 611–612
- OCR (optical character recognition), 321
- OpenRefine (data cleaning and transformation tool), 611, 613–615
- OSF (Open Science Framework), 16, 536, 544
- Phon (database software), 391, 395–397
- Phonological CorpusTools (PCT), 575, 576, 582
- Praat (speech analysis), 65, 196, 226–228, 584
- proprietary versus open, 340, 544, 561
- SayMore (language documentation tool), 287, 293, 296, 328
- Slack (collaboration tool), 540
- Suite of Automatic Linguistic Analysis Tools (SALAT), 414, 419
- Toolbox, 611
- Solar panels, portable, 330
- SOLIS (Sociolinguistics Server), 225, 230
- Source code, open, 432, 447, 450–451. *See also* Scripts
- Spectrograms, 30, 65, 423–424
- SPeech Across Dialects of English (SPADE), 195, 197–199, 205
- Speech analysis software, 196. *See also* Praat
- SpeechBox (Formerly the Online Speech/Corpora Archive and Analysis Resource (OSCAAR)), 241
- Speech database management systems, 196
 - Integrated Speech Corpus Analysis (ISCAN), 195, 196, 200–206
- Speech perception data, 565–566, 571–572
 - noise masking, 569
 - non-native listeners, 569
 - processing, 243
 - storage and sharing, 241, 244, 568, 570, 571–572
- Speech perception tasks, 239–240, 567, 569, 570
 - experimental scripts, 561
 - lexical decision tasks, 557, 559–560, 570
 - stimuli, 558–560
- Speech recognition technology, 195–196, 437–438, 444
 - back-off technique, 444
 - data quality, 440
 - language model, 443–444
 - signed language recognition, 463–464, 468
- Speech recognition tools, 195–196, 437–438
 - component models, 439
 - Elpis, 437–438, 439–445
 - ESPnet, 438, 445
 - Kaldi, 438, 445
- Spontaneous production data
 - signed language data, 465
 - spoken language data, 210–211, 218
- Spreadsheets
 - for coding, 216, 251–252, 457–458, 516–517, 613–615
 - for corpus linguistic annotation, 458
 - metadata spreadsheets, 251–253, 289–290, 329, 407
 - preventing data entry errors, 458, 459
 - used to track cognates, 333
- Statistical analyses
 - L2 learner data, 417
 - mixed-effects models, 413, 417, 457
 - natural language processing (NLP), 415–416
 - plots in R, 460
 - scripts, 447, 450–451, 562
- Storage of data. *See* Data storage
- Structural data, definition of, 50
- Subjective data, 492–494
- Survey of California and Other Indian Languages (California Language Archive), 134
- Sustainability
 - of file formats, 64–65, 75–77, 137, 260, 640 (*see also* File formats)
 - of research data, 63, 67–69, 94, 129, 137–138
- Swadesh vocabulary lists, 131, 346
- Synchronic analysis, 35
- Syntactic trees, 34, 499, 507
- Syntagmatic analyses, 33–34, 36
- Syntax, experimental, 531–533
 - counterbalancing, 534
 - data management, 531, 533–536
 - debriefing data, 535
 - demographic data, 535
- 3-2-1 rule (for data storage), 66, 106
- Tabular data, 35, 64, 75, 83
 - long-table formats, 347, 454–455, 534–535
 - multilingual word lists, 346–347
 - structuring strategy, 347
- TalkBank database system, 257, 391, 398
 - data citation, 399
- Tense, aspect, mood (TAM) systems, 609, 612
- Tenure process. *See* Review, promotion, and tenure (RPT) process
- TerraLing (searchable database collection), 617–619, 620
 - citation of, 626–627
 - native speaker or signer experts, 620, 625
- Text editors, 78
- Text files, 73–74, 75–78
- Textual data, 28–29
 - archival documentation, 315–317, 318–323
 - handwriting, 297, 319–320, 403, 406
 - Text Encoding Initiative (TEI) guidelines, 322
 - theoretical syntax, 528–529
 - transcription, 319–322
 - transcription reliability, 407

- Theoretical linguistics
- crosslinguistic data, 617, 621, 622, 627–628
 - theoretical syntactic data, 527–528
 - theoretical syntax, 524–526, 526–529
 - underinvestigated languages, 514, 526–527
- Theoretical syntax, 524
- data collection, 526–529, 618
 - database design, 524–526
 - elicited data, 527–528, 625
 - fieldwork workflow, 525, 526
 - Indigenous language data, 523
 - texts as data, 528–529
- Thomason, Sally, 4
- Tidy data principles, 559
- Time code format for data citation, 149
- Transcriber (annotation tool), 225
- Transcription
- automatic speech recognition (ASR), 438, 439, 444
 - autosegmentation, 293
 - based on indexed recordings, 251–252
 - CA (Conversation Analysis) research, 260–261
 - CHAT transcription (Codes for the Human Analysis of Transcripts), 370, 387, 406
 - of code-switching, 214
 - of contracted forms, 213
 - conventions, 188, 261, 293
 - of corpus interview data, 188, 212, 224–225
 - correction, 214
 - Cross-Linguistic Transcription Systems initiative, 347
 - forced alignment, 195–196, 225–226, 423–427
 - in field notebooks, 333
 - linguistically oriented versus forced alignment-oriented, 428
 - manual, 386–387
 - of non-overt (deleted) forms, 214, 215
 - OCR (optical character recognition) used for, 321
 - of oral and written corpora, 406
 - orthography used for, 213, 319–321, 441–442
 - phonetic, 388, 394
 - of signed language video data, 272, 369, 374–379, 465–467, 468
 - Speech Assessment Methods Phonetic Alphabet, 388
 - speech recognition tools, 437–438, 439–440
 - time-aligned, 251–253, 281–282, 296
 - tools for, 225, 281–282, 291–293, 296, 328, 437–438
 - and translation, 328, 332, 393–395
 - untrained forced alignment, 423–424, 427–429, 429–432
- Transfer agreements
- copyright transfer agreements, 122 (*see also* Copyright law)
 - data transfer agreements, 198 (*see also* Data sharing)
- Translations
- elicited, 528, 610
 - machine translation, 463, 486
 - as parallel data, 132, 499
 - parallel treebanks, 499, 503, 507
- Transliteration of data information content, 74
- Transparency, 5, 10–11, 12, 89, 598
- of methodology, 12, 14, 89, 531, 533
 - Peer Reviewer's Openness Initiative, 21
- Treebank data, 33, 499–500
- automatic parsing, 501, 503
 - citation of, 507–510
 - copyrighted material, 500
 - data collection, 500
 - INESS (Infrastructure for the Exploration of Syntax and Semantics), 501, 503, 505–507
 - metadata, 503
 - parallel treebanks, 499, 503, 507
 - parsing and annotation, 501
 - Penn Treebank, 499
 - sharing, 503
 - storage of, 502
- Tree diagrams, 34
- Tromsø Recommendations for Citation of Research Data in Linguistics, 5, 144
- Tromsø Repository of Language and Linguistics (TROLLing), 92, 94, 96, 448
- TU Delft Research Data Framework Policy, 95
- Twitter
- altmetrics, 159
 - as data source, 481–483, 484–485
- Typology, 38–40, 597–599
- AUTOTYP (typological database), 631–634, 634–638, 638–640
 - autotypology, 632–633
 - coding schemas, 623–624, 632, 634
 - comparative concepts, 597–598, 600
 - corpus-based, 597
 - data aggregation, 633–634, 638
 - database reference languages, 339
 - doculects, 598
 - GOLD ontology (General Ontology for Linguistic Description), 137
 - grammar-based, 597, 598
 - ImproType (Typological Database of Impersonals), 599–604, 604–606
 - “language” used as term in, 598
 - semantic map framework, 606
- UK General Data Protection Regulation (GDPR), 94, 96, 198–199, 398
- Underdescribed languages (Bantu language varieties), 609
- Underdocumented languages, 301, 514
- child language acquisition data (Cree), 392, 393–395, 397–398
 - data openness, 518
 - language and literacy resource materials, 301, 302
 - lexical data, 301, 302, 304
 - theoretical syntax, 523, 526–527
- Underinvestigated languages (Uzbek), 514, 518

- Underresourced languages, untrained forced alignment, 423–424, 424–427
- Unicode, 77
- Universal Numerical Fingerprint (UNF), 153, 165
- Unix tools, 83
- Unpublished data, citation of, 149
- Untrained forced alignment, 423–424. *See also* Forced alignment
- dictionaries for, 426
 - glyph management, 425–426, 430
 - underresourced languages, 423–424, 424–427
 - workflow, 424, 427–429, 429–432
- Uzbek
- grammar, 515, 516
 - language variation, 513–514, 517
 - sociocultural context, 513, 514, 518
- Variation, language, 215–216, 283
- Hawai'i Creole (Pidgin), 221
 - regional, 238
 - sign languages, 477
 - theoretical linguistic fieldwork, 528
 - Uzbek, 513–514, 517
- Variationist analyses, 215–216, 221
- sign languages, 477
- Version control, 75, 83–84
- ASL ID glosses, 377–378
 - DMPs (Data Management Plans), 113
 - dynamic versions, 147
 - Git (version control system), 84, 536, 540
 - in fieldwork setting, 329, 428
 - tools for, 84, 540, 563
 - transferring and updating files, 329
- Versioning. *See* Version control
- Video data, 28
- annotation of, 369, 374–375, 378–379
 - anonymization and deidentification of, 254, 262, 270, 398
 - collection of, 250, 259, 393, 464–465
 - compatibility with annotation software, 273
 - ethical considerations for, 259, 464–465, 472, 473–476
 - metadata, 251–253, 260, 331, 374, 378–379
 - multiple camera, 472
 - quality of, 464, 465
 - segmentation and transcription of, 251–253, 331, 393–395
 - signed languages, 370–372, 374–379, 464–465, 472–475
 - of social interactions, 250, 257
 - storage and backup of, 251, 372–374
- Visibility of scholarship, 90, 161–162, 167
- Vowels in America (VIA) project, 237, 238, 244
- Wansink, Brian, 17
- Wax cylinder recordings, 137, 316
- Web companion course (open access), 6–7
- Word lists, 38
- concept sets, 347
 - elicitation using, 131, 328
 - for historical reconstruction, 333
 - lexicostatistics, 348
 - limitations of, 304–305
 - multilingual, 345, 346–347
 - Swadesh lists, 131, 346
 - uses of, 339
- Word recognition (naming) task, 416
- Workflow. *See also names of individual subfields of linguistics*
- asynchronous for teams, 432, 563
 - automatic speech recognition (ASR), 439–440
 - corpus linguistics, 454–456, 456–458, 459, 460
 - data workflow, 62, 66, 108, 110 (*see also* Data management)
 - documentary linguistics, 282, 288, 331–332
 - experimental research, 540, 541–542, 543–544, 563
 - lexicography, 304, 307–309, 310, 311–312
 - linguistic fieldwork, 62, 64, 327–331
 - metadata documentation, 108–109, 110, 329, 331 (*see also* Metadata)
 - pipelines and protocols, 340
 - processing and storage of recordings, 251–253, 329 (*see also* Audio recordings; Video data)
 - research workflow, 62, 66, 108, 110, 126
 - untrained forced alignment, 424, 427–429, 429–432
- Working data, versus raw data, 106
- World Atlas of Language Structures (WALS), 618, 638
- language codes, 632
- Writing systems, 132–133. *See also* Orthography
- for archiving data, 519
 - ideographic writing, 132
- Zotero, 152

This is a section of [doi:10.7551/mitpress/12200.001.0001](https://doi.org/10.7551/mitpress/12200.001.0001)

The Open Handbook of Linguistic Data Management

Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister

Citation:

The Open Handbook of Linguistic Data Management

Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister

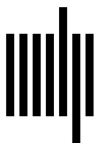
DOI: 10.7551/mitpress/12200.001.0001

ISBN (electronic): 9780262366076

Publisher: The MIT Press

Published: 2022

The open access edition of this book was made possible by generous funding and support from the authors



The MIT Press

© 2021 The Massachusetts Institute of Technology

This work is subject to a Creative Commons CC-BY-NC license. Subject to such license, all rights are reserved.



This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Berez-Kroeker, Andrea L., editor. | McDonnell, Bradley James, editor. | Koller, Eve, editor. | Collister, Lauren B., editor.

Title: The open handbook of linguistic data management / edited by Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller and Lauren B. Collister.

Description: Cambridge, Massachusetts : The MIT Press, [2021] | Series: Open handbooks in linguistics series | Includes bibliographical references and index.

Identifiers: LCCN 2020044363 | ISBN 9780262045261 (hardcover)

Subjects: LCSH: Computational linguistics. | Natural language processing (Computer science) | Data mining.

Classification: LCC P98 .O64 2021 | DDC 410.285—dc23

LC record available at <https://lcn.loc.gov/2020044363>